Variational Hidden Conditional Random Fields with Coupled Dirichlet Process Mixtures

Konstantinos Bousmalis[1], Stefanos Zafeiriou[1], Louis–Philippe Morency[2], Maja Pantic[1], and Zoubin Ghahramani[3]

[1] Imperial College London, SW7 2AZ, UK
{k.bousmalis,s.zafeiriou,m.pantic}@imperial.ac.uk
[2] University of Southern California, Playa Vista, CA 90094 USA
morency@ict.usc.edu
[3] University of Cambridge, CB2 1PZ, UK
zoubin@eng.cam.ac.uk

Abstract. Hidden Conditional Random Fields (HCRFs) are discriminative latent variable models which have been shown to successfully learn the hidden structure of a given classification problem. An infinite HCRF is an HCRF with a countably infinite number of hidden states, which rids us not only of the necessity to specify a priori a fixed number of hidden states available but also of the problem of overfitting. Markov chain Monte Carlo (MCMC) sampling algorithms are often employed for inference in such models. However, convergence of such algorithms is rather difficult to verify, and as the complexity of the task at hand increases, the computational cost of such algorithms often becomes prohibitive. These limitations can be overcome by variational techniques. In this paper, we present a generalized framework for infinite HCRF models, and a novel variational inference approach on a model based on coupled Dirichlet Process Mixtures, the HCRF–DPM. We show that the variational HCRF–DPM is able to converge to a correct number of represented hidden states, and performs as well as the best parametric HCRFs —chosen via cross–validation— for the difficult tasks of recognizing instances of agreement, disagreement, and pain in audiovisual sequences.

Keywords: nonparametric models, discriminative models, hidden conditional random fields, dirichlet processes, variational inference.

1 Introduction

Over the past decade, nonparametric methods have been successfully applied to many existing graphical models, allowing them to grow the number of latent states as necessary to fit the data [1–6]. Infinite HCRFs were first presented in [7] and since exact inference for such models with an infinite number of parameters is intractable, inference was based on a Markov chain Monte Carlo (MCMC) sampling algorithm. Although MCMC algorithms have been successfully applied on numerous applications, they have some significant drawbacks: they are notoriously slow to converge, it is hard to verify their convergence,

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part II, LNAI 8189, pp. 531–547, 2013.

and they often don't scale well to larger datasets and higher model complexity. Moreover, the model presented in [7] is not readily able to handle continuous input features.

In this work, we consider a deterministic alternative to MCMC sampling for infinite HCRFs with a variational inference [8] approach. Variational inference will allow us to converge faster, verify convergence and scale without a prohibitive computational cost. The model we present in this paper allows a countably infinite number of shared, among labels, hidden states via the use of multiple Dirichlet Process Mixtures (DPMs). Specifically, we present a novel mean field variational approach that uses DPM constructions in the model potentials to allow for the representation of a potentially infinite number of hidden states. Furthermore, we show that our model, the HCRF–DPM, is a generalization of the model presented in [7] and is able to handle continuous features naturally.

In the following section, we consicely present the theoretical background necessary to understand this paper. We present in Section 3 our variational HCRF–DPM model. Finally, we evaluate our model in Section 4.2, and conclude in Section 5.

2 Theoretical Background

The HCRF–DPM, like many other infinite models, relies on DPMs. We present in this section a brief introduction to Dirichlet Processes and Hidden Conditional Random Fields.

2.1 Dirichlet Processes

A Dirichlet Process (DP) is a distribution of distributions, parameterized by a scaling parameter α and a probability measure Ξ. The latter is the basis around which the distributions $G \sim \mathrm{DP}(\alpha, \Xi)$ are drawn, with variability governed by the α parameter. [9] presented the so–called "stick–breaking" construction for DPs, which is based on random variables $(\beta'_k)_{k=1}^{\infty}$ and $(h_k)_{k=1}^{\infty}$, where $\beta'_k|\alpha, \Xi \sim Beta(1, \alpha)$ and $h_k|\alpha, \Xi \sim \Xi$:

$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l) \qquad G = \sum_{k=1}^{\infty} \beta_k \delta_{h_k} \ , \tag{1}$$

where δ is the Dirac delta function. By letting $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ we abbreviate this construction as $\boldsymbol{\beta}|\alpha \ \sim GEM(\alpha)$. A Dirichlet Process Mixture (DPM) model is a hierarchical Bayesian model that uses a DP as a nonparametric prior:

$$G|\alpha, \Xi \sim \mathrm{DP}(\alpha, \Xi), \quad c_t \mid G \sim G, \quad s_t \ \sim p(s_t|c_t) \ , \tag{2}$$

where $(s_t)_{t=1}^{T}$ is a dataset of size T, governed by a distribution conditioned on $(c_t)_{t=1}^{T}$, auxiliary index variables that get assigned each to one of the clusters $(h_k)_{k=1}^{\infty}$. As new datapoints are drawn, the number of components in this mixture model grows. In the model we present in this paper, as we explain later, we employ a number of DP priors coupled together at the data generation level, i.e. s_t above is a function of auxiliary index variables drawn from all different DPs.

2.2 Finite Hidden Conditional Random Fields

HCRFs —discriminative undirected models that contain hidden states— were first presented in [10] and used to capture temporal dependencies across frames and recognize different gesture classes. They did so successfully by learning a state distribution among the different gesture classes in a discriminative manner, allowing them to not only uncover the distinctive configurations that uniquely identify each class, but also to learn a shared common structure among the classes. Conditional Random Fields and HCRFs can be defined in arbitrary graph structures but in our paper, driven by our application field, we assume data to be sequences that correspond to undirected chains. Our work, however, can be readily applied to tree–structured models.

We represent T observations as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$. Each observation at time $t \in \{1, \ldots, T\}$ is represented by a feature vector $\mathbf{f}_t \in \Re^d$, where d is the number of features, that can include any features of the observation sequence. We wish to learn a mapping between observation sequence $\mathbf{X}$ and class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of available labels. The HCRF does so by estimating the conditional joint distribution over a sequence of latent variables $\mathbf{s} = [s_1, s_2, \ldots, s_T]$, each of which is assigned to a hidden state $h_k \in \mathcal{H}$, and a label y, given $\mathbf{X}$. One of the main representational power of HCRFs is that the latent variables can depend on arbitrary features of the observation sequence. This allows us to model long–range contextual dependencies: s_t, the latent variable at time t, can depend on observations that happened earlier or later than t. An HCRF models the conditional probability of a class label given an observation sequence by:

$$p(y \mid \mathbf{X}, \boldsymbol{\theta}) = \sum_{\mathbf{s}} p(y, \mathbf{s} \mid \mathbf{X}, \boldsymbol{\theta}) = \frac{\sum_{\mathbf{s}} \mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta})}{\sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \mathbf{X}, \boldsymbol{\theta})} \ . \tag{3}$$

The potential function $\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \in \Re$ is parameterized by $\boldsymbol{\theta}$, which measures the compatibility between a label y, a sequence of observations $\mathbf{X}$ and a configuration of the latent variables $\mathbf{s}$. The model is discriminative because it doesn't model a joint distribution that includes input $\mathbf{X}$, but it only models the distribution of a label y conditioned on $\mathbf{X}$. The graph of a linear–chain HCRF is a chain where each node corresponds to a latent variable s_t at time t. For such a model, the potential function is usually defined as:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \ = \ \exp\left\{ \sum_{t=1}^{T} \sum_{i=1}^{d} \theta_x(s_t, i) f_t(i) + \theta_y(s_t, y) + \sum_{t=2}^{T} \theta_e(s_t, s_{t-1}, y) \right\} \tag{4}$$

In this paper, we use the notation $\theta_x(h_k, i)$ to refer to the weight that measures the compatibility between the feature indexed by i and state $h_k \in \mathcal{H}$. Similarly, $\theta_y(h_k, y)$ stand for weights that correspond to class y and state h_k, whereas $\theta_e(h_k, h', y)$ measure the compatibility of y with a transition from h' to h_k.

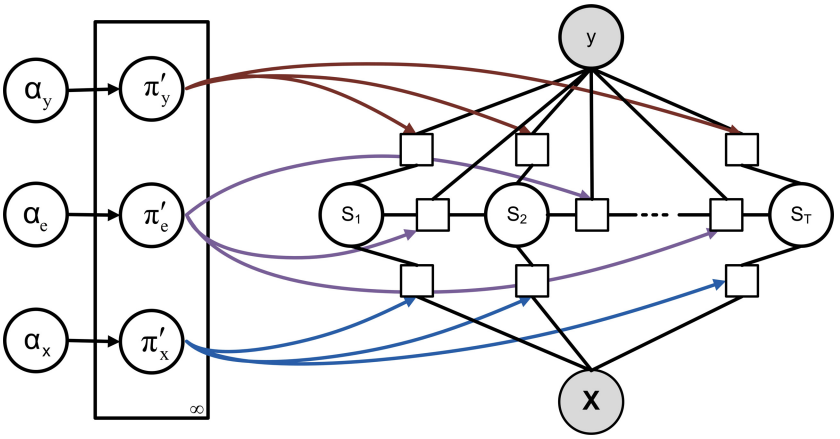


Fig. 1. Factor graph representation of our HCRF–DPM

3 Hidden Conditional Random Fields with Coupled Dirichlet Process Mixtures

For an infinite HCRF we allow an unbounded number of potential hidden states in $\mathcal{H}$. This becomes possible, by introducing random variables $\{\pi_x(h_k|i)\}_{k=1}^{\infty}$, $\{\pi_y(h_k|y)\}_{k=1}^{\infty}$, $\{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|}$ for an observation feature indexed by i, label y, and an assignment $s_{t-1} = h_a$. These new random variables are drawn by distinct processes that are able to model such quantities and are subsequently incorporated in the node and edge potentials of our HCRF. We present in this paper the HCRF–DPM, a model that uses DPMs to define these random quantities (see its factor graph representation in Fig. 1). These variables, even though drawn by distinct processes, are coupled together by a common latent variable assignment in our graphical model. We redefine our potential function $\mathcal{F}$ from (4) as follows:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) = \exp\left\{\sum_{t=1}^{T}\sum_{i=1}^{d}\theta_x(s_t, i)f_t(i)\log\pi_x(s_t|i) + \theta_y(s_t, y)\log\pi_y(s_t|y) + \sum_{t=2}^{T}\theta_e(s_t, s_{t-1}, y)\log\pi_e(s_t, y|s_{t-1})\right\}. \quad (5)$$

We assume that random variables $\{\pi_x(h_k|i)\}_{k=1}^{\infty}$, $\{\pi_y(h_k|y)\}_{k=1}^{\infty}$, $\{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|}$ are between 0 and 1. These are in effect the quantities that will allow the model to 'select' an appropriate number of useful hidden states for a given classification task. $\mathbf{f}_t$ are nonnegative features extracted from the observation sequence $\mathbf{X}$ and, as before, they can include arbitrary features of the input. We assume that $\boldsymbol{\theta}$ are nonnegative parameters and, as in (4), they model the relationships between hidden states and features ($\boldsymbol{\theta}_x$),

labels ($\boldsymbol{\theta}_y$) and transitions ($\boldsymbol{\theta}_e$). These nonnegativity constraints for $\boldsymbol{\theta}$ and $\mathbf{f}$ are essential in this model, since the π-quantities are random variables and influence the probabilities of the hidden states: a negative parameter or feature would make an otherwise improbable state very likely to be chosen. Moreover, these constraints ensure compliance with the positivity constraints of our variational parameter updates (25)-(30), as we shall see later in this section. Finally, it is important to note that the positivity of $\boldsymbol{\theta}$ is not theoretically restrictive for our model due to the HCRF normalization factor $\frac{1}{Z(\mathbf{X})}$ in (3) where

$$Z(\mathbf{X}) = \sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \mathbf{X}, \boldsymbol{\theta}) \ . \tag{6}$$

The HCRF–DPM model is an infinite HCRF where the quantities $\{\pi_x(h_k|i)\}_{k=1}^{\infty}$, $\{\pi_y(h_k|y)\}_{k=1}^{\infty}$, $\{\pi_e(h_k, y|h_a)\}_{k=1,y=1}^{\infty,|\mathcal{Y}|}$ in (5) are driven by coupled DPMs. It is important to understand that for the DPMs driving the π_e quantities in the edge features, h_k and y are treated as a single random variable –their product– $\omega_\mu = \{h_k, y\}$ that effectively has a state–space of size $|\mathcal{Y}| \times |\mathcal{H}|$, still an infinite number. According to the stick–breaking properties of DPs, we construct $\boldsymbol{\pi} = \{\boldsymbol{\pi}_x, \boldsymbol{\pi}_y, \boldsymbol{\pi}_e\}$ conditioned on a new set of random variables $\boldsymbol{\pi}' = \{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e\}$ that follow $Beta$ distributions:

$$\pi'_x(h_k|i) \sim Beta(1, \alpha_x), \qquad \pi_x(h_k|i) = \pi'_x(h_k|i) \prod_{j=1}^{k-1} (1 - \pi'_x(h_j|i)) \tag{7}$$

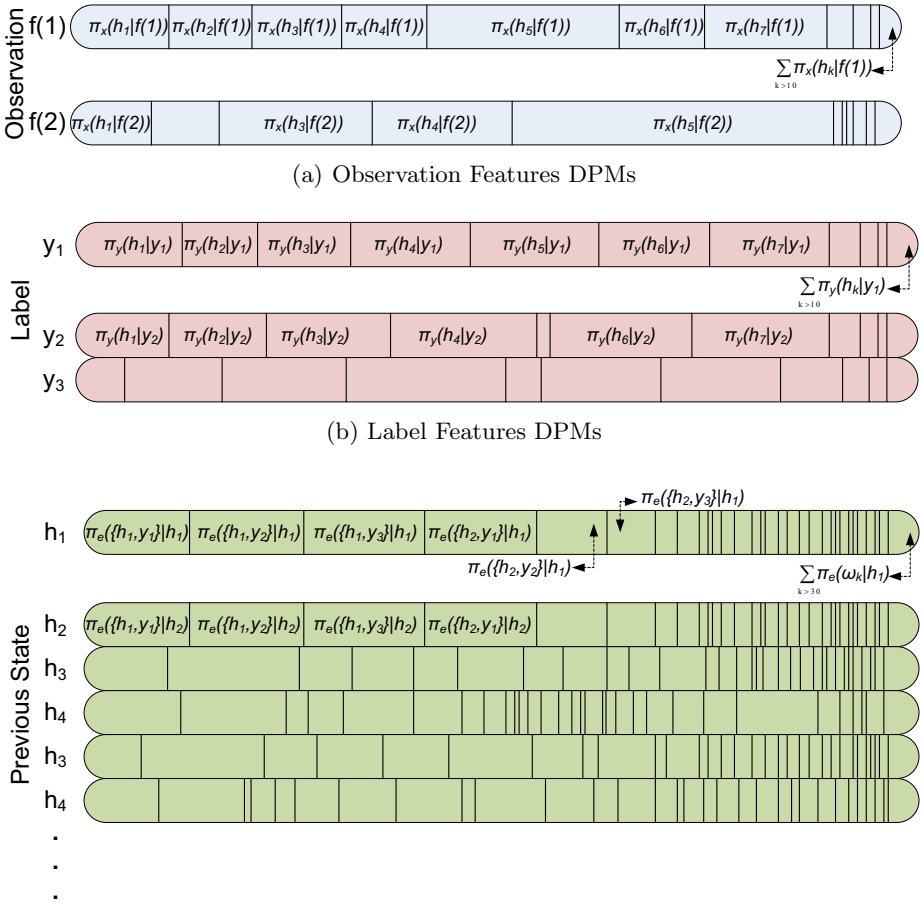
$$\pi'_y(h_k|y) \sim Beta(1, \alpha_y), \qquad \pi_y(h_k|y) = \pi'_y(h_k|y) \prod_{j=1}^{k-1} \left(1 - \pi'_y(h_j|y)\right) \tag{8}$$

$$\pi'_e(\omega_\mu|h_a) \sim Beta(1, \alpha_e), \qquad \pi_e(\omega_\mu|h_a) = \pi'_e(\omega_\mu|h_a) \prod_{j=1}^{\mu-1} (1 - \pi'_e(\omega_\mu|h_a)) \tag{9}$$

This process can be made clearer by examining Fig. 2, where we visualize the stick breaking construction of an HCRF–DPM model with 2 observation features, 3 labels, and 10 'important' hidden states. The π_e-sticks have an important —for the implementation of our model— difference to the π_x and π_y–sticks in that the hidden states are intertwined with the labels, with each stick piece representing an ω–state. This means there are $|\mathcal{Y}|$ such states corresponding to one h–state. This becomes particularly important later on when we calculate our variational updates.

By using (5) the sequence of latent variables $\mathbf{s} = \{s_1, ...s_T\}$ can then be generated by the following process:

1. Draw $\pi'_x|\alpha_x \sim Beta(1, \alpha_x)$, $\pi'_y|\alpha_y \sim Beta(1, \alpha_y)$, $\pi'_e|\alpha_e \sim Beta(1, \alpha_e)$
2. Calculate $\boldsymbol{\pi}$ from (7)-(9). Note that this will only need to be calculated for a finite number of hidden states, due to our variational approximation.
3. For the t^{th} latent variable, using (5) we draw



(a) Observation Features DPMs

(b) Label Features DPMs

(c) Transition Features DPMs

Fig. 2. Visualization of the π-'sticks' used to construct the infinite states in our HCRF–DPM. The fictitious model presented here has 2 observation features $f(1), f(2)$, 3 labels y_1, y_2, y_3 and fewer than 10 important hidden states $h_1, h_2, h_3 \ldots$. Each 'stick' sums up to 1, and the last piece always represents the sum of the lengths that correspond to all hidden states after the 10^{th} state. Notice that for the π_e-'sticks' this corresponds to 30 ω–states. For example $\pi_e(h_1, y_3|h_2)$ controls the probability of transitioning from h_2 to h_1 in a sequence with label y_3. See text for more details.

$$s_t|\{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e, s_{t-1}, y, \mathbf{X}\} \sim Mult\Bigg(\exp\Big\{ \sum_{i=1}^{d} \theta_x(s_t, i) f_t(i) \log \pi_x(s_t|i) + \theta_y(s_t, y) \log \pi_y(s_t|y) + \theta_e(s_t, s_{t-1}, y) \log \pi_e(\{s_t, y\}|s_{t-1}) \Big\} \Bigg) \tag{10}$$

Rather than expressing the model in terms of $\boldsymbol{\pi}$, we use $\boldsymbol{\pi}' = \{\boldsymbol{\pi}'_x, \boldsymbol{\pi}'_y, \boldsymbol{\pi}'_e\}$ resulting in the folowing joint distribution that describes the HCRF–DPM:

$$p(y, \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}, \theta) \;=\; p(y, \mathbf{s} \mid \boldsymbol{\pi}', \mathbf{X}, \theta) p(\boldsymbol{\pi}'_x) p(\boldsymbol{\pi}'_y) p(\boldsymbol{\pi}'_e) \tag{11}$$

with

$$p(y, \mathbf{s} \mid \boldsymbol{\pi}', \mathbf{X}, \theta) \;=\; \frac{1}{Z(\mathbf{X})} \mathcal{F}(y, \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}, \boldsymbol{\theta}) \tag{12}$$

where $Z(\mathbf{X}) = \sum_{y' \in \mathcal{Y}, \mathbf{s}} \mathcal{F}(y', \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}, \boldsymbol{\theta})$. We assume independence of all π' variables above, so for example $p(\boldsymbol{\pi}'_x) = \prod_{k=1}^{\infty} \prod_{i=1}^{d} \pi'_x(h_k|i)$.

Comparison with Previous Work. It is important at this stage to compare our model described by (5) with the MCMC model (IHCRF–MCMC) presented in [7]. The latter work defined potentials for each of the relationships between hidden states and features, labels and transitions and the potential function $\mathcal{F}$ as their product along the model chain:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}) = \mathcal{F}_x(\mathbf{s}, \mathbf{X}) \mathcal{F}_y(y, \mathbf{s}) \mathcal{F}_e(y, \mathbf{s}) \tag{13}$$

$$\mathcal{F}_x(\mathbf{s}, \mathbf{X}) = \prod_{t=1}^{T} \prod_{i=1}^{d} \pi_x(s_t|i)^{f_t(i)} \tag{14}$$

$$\mathcal{F}_y(y, \mathbf{s}) = \prod_{t=1}^{T} \pi_y(s_t|y) \tag{15}$$

$$\mathcal{F}_e(y, \mathbf{s}) = \prod_{t=2}^{T} \pi_e(y, s_t|s_{t-1}) \tag{16}$$

The quantities $\boldsymbol{\pi}_x, \boldsymbol{\pi}_y, \boldsymbol{\pi}_e$ above are conceptually the same as in our model, except for the fact that in [7] they have Hierarchical Dirichlet Process (HDP) priors instead of DP priors, as we do in this paper.[1]

[1] Using HDP priors allows separate DPMs to be linked together via an identical base probabilistic measure, which is itself a DP. It would be interesting to use such priors for our model, but we were able to obtain satisfactory results without introducing higher complexity and additional hyperparameters into the variational model we experimented with. Notice that our model allows for such flexibility: using HDP priors would simply change the updates for our variational coordinate descent algorithm.

The potential function (13) above can be rewritten as follows:

$$\mathcal{F}(y, \mathbf{s}, \mathbf{X}) \;=\; \exp\left\{\sum_{t=1}^{T}\sum_{i=1}^{d} f_t(i)\log\pi_x(s_t|i) + \log\pi_y(s_t|y) + \sum_{t=2}^{T}\log\pi_e(s_t, y|s_{t-1})\right\} \tag{17}$$

A comparison between (17) and (5) makes it clear that our model is a generalization of the model presented in [7], which assumes, according to our framework, that $\boldsymbol{\theta}$-parameters are set to 1. The introduction of these parameters is not redundant, but allows for a more powerful and flexible models. Also, when dealing with classification problems involving continuous observation features using (5) for the potential function of an infinite HCRF is more suitable than (17), as we show in the experimental section. In those cases it is known that θ–parameters are of particular importance as they are able to capture the scaling of each input feature. The former model is not guaranteed to perform well unless some non–trivial normalization is applied on the observation features.

3.1 Variational Inference for the HCRF–DPM

Since inference on our model (11) is intractable, we need to approximate the marginal probabilities along the chain of our graphical model, and the π–quantities in (5). We shall do so with a mean–field variational inference approach. We use the following approximation for the joint distribution of our model:

$$q(y, \mathbf{s}, \boldsymbol{\pi}', \mathbf{X}) = q(y, \mathbf{s}|\mathbf{X})q(\boldsymbol{\pi}'_x)q(\boldsymbol{\pi}'_y)q(\boldsymbol{\pi}'_e) \tag{18}$$

where,

$$\begin{aligned} q(y, \mathbf{s}|\mathbf{X}) &= q(y, s_1|\mathbf{X})\prod_{t=2}^{T} q(y, s_t|s_{t-1}, \mathbf{X}) \\ &= \prod_{i=1}^{d} q(s_1|i)q(s_1|y)\prod_{t=2}^{T}\prod_{i=1}^{d} q(s_t|i)q(s_t|y)q(s_t, y|s_{t-1}) \;. \end{aligned} \tag{19}$$

Each individual approximate $q(\pi'_x), q(\pi'_y), q(\pi'_e)$ follows a *Beta* distribution with variational parameters $\boldsymbol{\tau}_x, \boldsymbol{\tau}_y, \boldsymbol{\tau}_e$ respectively. Explicitly, for features indexed by i, labels indexed by y, and hidden states indexed by k, k':

$$q(\pi'_x(h_k|i)) = Beta\left(\tau_{x,1}(k,i), \tau_{x,2}(k,i)\right) \tag{20}$$

$$q(\pi'_y(h_k|y)) = Beta\left(\tau_{y,1}(k,y), \tau_{y,2}(k,y)\right) \tag{21}$$

$$q(\pi'_e(y, h_k|h_{k'})) = Beta\left(\tau_{e,1}(y,k,k'), \tau_{e,2}(y,k,k')\right) \tag{22}$$

We approximate all $\boldsymbol{\pi}$ variables by employing a truncated stick–breaking representation which approximates the infinite number of hidden states with a finite number L [11]. This is the crux of our variational approach, and it effectively means that we set a truncation threshold L, above which the above quantities

are set to 0: $\forall k > L$, $q(\pi'_x(h_k|i)) = 0$, $q(\pi'_y(h_k|y)) = 0$, $q(\pi'_e(y, h_k|h_{k'})) = 0$. Note that using this approximation is statistically rather different from using a finite model: an HCRF–DPM simply approximates the infinite number of states and will still reduce the number of useful hidden states to something smaller than L. It is finally important to stress that by constraining our θ–parameters and observation features to be positive, we effectively make the number of the θ–parameters that matter finite: changing a θ–parameter associated with a hidden state $k > L$ will not change our model.

3.2 Model Training

A trained variational HCRF–DPM model is defined as the set of optimal parameters $\boldsymbol{\theta}^*$ and optimal variational parameters $\boldsymbol{\tau}^*$. In this work we obtain these with a training algorithm that can be divided in two distinct phases: *(i)* the optimization of our variational paramaters through a coordinate descent algorithm using the updates derived below and *(ii)* the optimization of parameters $\boldsymbol{\theta}$ through a gradient descent method. Although it would be possible to have a fully Bayesian model with $\boldsymbol{\theta}$ being random variables in our model, inference would become more difficult. Moreover, having a single value for our $\boldsymbol{\theta}$ parameters is good for model interpretability and makes the application of a trained model to test data much easier.

Phase 1: Optimization of Variational Parameters $\boldsymbol{\tau}$. Now that we have defined an approximate model distribution in (18), we can approximate the necessary quantities for our inference. These approximations, as one can see later in this section, depend solely on our variational parameters $\boldsymbol{\tau}$. We calculate those by minimizing the Kullback-Liebler divergence (KL) between approximate and actual joint distributions of our model, (11) and (18), using a coordinate descent algorithm:

$$\begin{aligned} KL[q||p] = \log Z(\mathbf{X}) &- \langle \log \mathcal{F}(y, \mathbf{x}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}') \rangle_{q(y,\mathbf{s},\boldsymbol{\pi}'|\mathbf{X})} \\ &+ \langle \log q(y, \mathbf{s}|\mathbf{X}) q(\boldsymbol{\pi}') \rangle_{q(y,\mathbf{s},\boldsymbol{\pi}'|\mathbf{X})} \end{aligned} \tag{23}$$

where $\langle \cdot \rangle_q$ is the expectation of $\cdot$ with respect to q. Thus, the energy of the configuration of our random variables $y, \mathbf{s}$, and $\boldsymbol{\pi}'$ is $\log \mathcal{F}(y, \mathbf{x}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}')$ and the free energy of the variational distribution:

$$\mathcal{L}(q) = - \langle \log \mathcal{F}(y, \mathbf{x}, \boldsymbol{\pi}', \mathbf{X}) p(\boldsymbol{\pi}') \rangle_{q(y,\mathbf{s},\boldsymbol{\pi}'|\mathbf{X})} + \langle \log q(y, \mathbf{s}|\mathbf{X}) q(\boldsymbol{\pi}') \rangle_{q(y,\mathbf{s},\boldsymbol{\pi}'|\mathbf{X})} \tag{24}$$

Since $\log Z(\mathbf{X})$ is constant for a given observation sequence, minimizing the free energy $\mathcal{L}(q)$ minimizes the KL divergence.

We will obtain the variational updates for the two groups of latent variables $q(y, \mathbf{s}|\mathbf{X})$ and $q(\boldsymbol{\pi}')$ by setting the partial derivative with respect to each group of $\mathcal{L}(q)$ to 0 and solving for the approximate distribution of each group of latent variables. The updates for the *Beta* parameters of $q(\boldsymbol{\pi}')$ from (20)-(22) are:

$$\tau_{x,1}(k,i) = \sum_t f_t[i]\theta_x(k,i)q(s_t = h_k|i) + 1 \tag{25}$$

$$\tau_{x,2}(k,i) = \sum_t f_t[i]\theta_x(k,i)q(s_t > h_k|i) + \alpha_x \tag{26}$$

$$\tau_{y,1}(y,i) = \sum_t \theta_y(k,y)q(s_t = h_k|y) + 1 \tag{27}$$

$$\tau_{y,2}(y,i) = \sum_t \theta_y(k,i)q(s_t > h_k|y) + \alpha_y \tag{28}$$

$$\tau_{e,1}(y,k,k') = \sum_t \theta_e(k,k',y)q(s_t = h_k, y, s_{t-1} = h_{k'}) + 1 \tag{29}$$

$$\tau_{e,2}(y,k,k') = \sum_t \theta_e(k,k',y)q(s_t > h_k, y, s_{t-1} = h_{k'}) + \alpha_e \tag{30}$$

Quantities $q(s_t = h_k|i)$, $q(s_t = h_k|y)$, and $q(s_t = h_k, y, s_{t-1} = h_{k'})$ can be obtained by the forward–backward algorithm. The latter requires only conditional approximate likelihoods $q(s_t = h_k|i, y, h_{k'})$, which can be also be calculated by setting the derivative of $\mathcal{L}(q)$ to zero:

$$\begin{aligned} q(s_t = h_k|i, y, h_{k'}) \propto \exp\Bigg\{ & \\ & f_t(i)\theta_x(k,i)\left(\langle \log \pi'_x(s_t = h_k|i)\rangle_{q(\boldsymbol{\pi}')} + \sum_{j=k+1}^{L} \langle \log(1 - \pi'_x(s_t = h_j|i))\rangle_{q(\boldsymbol{\pi}')}\right) \\ & \theta_y(k,y)\left(\left\langle \log \pi'_y(s_t = h_k|y)\right\rangle_{q(\boldsymbol{\pi}')} + \sum_{j=k+1}^{L} \left\langle \log(1 - \pi'_y(s_t = h_j|y))\right\rangle_{q(\boldsymbol{\pi}')}\right) \\ & \theta_e(k,k',y)\Bigg(\langle \log \pi'_e(s_t = h_k, y|s_{t-1} = h_{k'})\rangle_{q(\boldsymbol{\pi}')} + \\ & \qquad \sum_{j=k+1}^{L} \langle \log(1 - \pi'_e(s_t = h_j, y|s_{t-1} = h_{k'}))\rangle_{q(\boldsymbol{\pi}')}\Bigg)\Bigg\} \end{aligned} \tag{31}$$

Since all $\boldsymbol{\pi}'$ follow a Beta distribution, the expectations above are known.

Phase 2: Optimization of Parameters $\boldsymbol{\theta}$. We find our optimal parameters $\boldsymbol{\theta}^* = \arg\max \log p(y|\mathbf{X}, \boldsymbol{\theta})$ based on a training set by using a common HCRF quasi–Newton gradient descent method (LBFGS), which requires the gradient of the log–likelihood with respect to each parameter. These gradients for our model are:

$$\frac{\partial \log p(y|\mathbf{X},\boldsymbol{\theta})}{\partial \theta_x(k,i)} = \sum_t p(s_t = h_k|y,\mathbf{X},\boldsymbol{\theta}) f_t(i) \log \pi_x(h_k|i) - \sum_{y' \in \mathcal{Y},t} p(s_t = h_k, y'|\mathbf{X},\boldsymbol{\theta}) f_t(i) \log \pi_x(h_k|i) \quad (32)$$

$$\frac{\partial \log p(y|\mathbf{X},\boldsymbol{\theta})}{\partial \theta_y(k,y)} = \sum_t p(s_t = h_k|y,\mathbf{X},\boldsymbol{\theta}) \log \pi_y(h_k|y) - \sum_{y' \in \mathcal{Y},t} p(s_t = h_k, y'|\mathbf{X},\boldsymbol{\theta}) \log \pi_y(h_k|y) \quad (33)$$

$$\frac{\partial \log p(y|\mathbf{X},\boldsymbol{\theta})}{\partial \theta_e(k,k',y)} = \sum_t p(s_t = h_k, s_{t-1} = h_{k'}|y,\mathbf{X},\boldsymbol{\theta}) \log \pi_e(h_k, y|h_{k'}) - \sum_{y' \in \mathcal{Y},t} p(s_t = h_k, s_{t-1} = h_{k'}, y'|\mathbf{X},\boldsymbol{\theta}) \log \pi_e(h_k, y|h_{k'}) \quad (34)$$

We make this gradient descent tractable by using the variational approximations for the intractable quantities in the above equations. However, there is a significant difference with other CRF and HCRF models that use such techniques to find optimal parameters: we are constrained to only positive θ-parameters Since we are using a quasi–Newton method with Armijo backtracking line search, we can use the gradient projection method of [12, 13] to enforce this constrain. Finally, it is important to stress here that, although our model includes parameters that are not treated probabilistically, we have not seen signs of overfitting in our experiments (see Fig. 4).

4 Experimental Results

4.1 Performance on a Synthetic Dataset with Continuous Features

In an effort to demonstrate the ability of our HCRF–DPM to model sequences with continuous features correctly, we created a synthetic dataset, on which we compared its performance to that of the IHCRF–MCMC model [7]. The simple dataset was generated by two HMMs, with 4 Gaussian hidden states each. Two of the states were shared between the two HMMs, resulting in a total of 6 unique hidden states, out of a total of 8 for the two labels.

We trained 10 randomly initialized models of the finite HCRF, IHCRF–MCMC and HCRF–DPM on 100 training sequences and chose in each case the best one based on their performance on an evaluation set of 100 different sequences. The performance of the models was finally evaluated by comparing the F1 measure achieved on a test set of 100 other sequences. All sets had an equal number of samples from each label. The IHCRF–MCMC model was unable to solve this simple two–label sequence classfication problem with continuous-only input features: it consistently selected Label 1. On the other hand, the finite HCRF and the new HCRF–DPM model were successful in achieving a perfect F1 score of 100% on the test set (see Table 1).

4.2 Application to the Audiovisual Analysis of Human Behavior

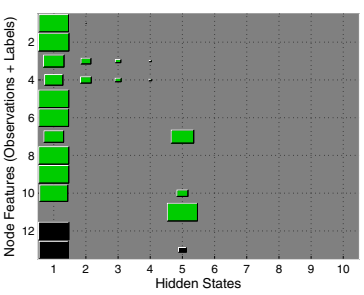
The problem of automatically classifying episodes of high–level emotional states, such as pain, agreement and disagreement, based on nonverbal cues in audiovisual sequences of spontaneous human behavior is rather complex [14]. Although humans are particularly good at interpreting such states, automated systems perform rather poorly. Infinite models are particularly attractive for modeling human behavior as we usually cannot have a solid intuition regarding the number of hidden states in such applications. Furthermore, it opens up the way of analyzing the hidden states these models converge to, which might provide social scientists with valuable information regarding the temporal interaction of groups of behavioral cues that are different or shared in these behaviors. We therefore decided to evaluate our novel approach on behavior analysis and specifically the recognition of agreement, disagreement and pain in recordings of spontaneous human behavior. We expected that our HCRF–DPM models would find a good number of shared hidden states and perform at least as well as the best cross–validated finite HCRF and IHCRF–MCMC models.

In this work we used an audiovisual dataset of spontaneous agreement and disagreement and a visual dataset of pain to evaluate the performance of the proposed model on four classification problems: *(1)* ADA2, agreement and disagreement recognition with two labels (agreement vs. disagreement); *(2)* ADA3, agreement and disagreement recognition with three labels (agreement vs. disagreement vs. neutral); *(3)* PAIN2, pain recognition with two labels (strong pain vs. no pain); and *(4)* PAIN3, pain recognition with three labels (strong pain vs. moderate pain vs. no pain). We show that *(1)* our model is capable of finding a good number of useful states; and *(2)* HCRF–DPMs perform better than the best performing finite HCRF and IHCRF–MCMC models in all of these problems with the exception of ADA3, where the performance of the HCRF–DPM is similar to that of the finite model.

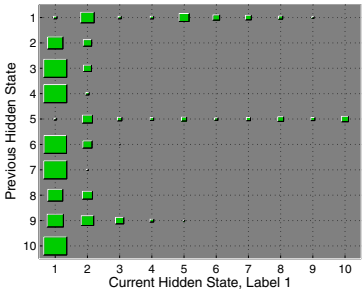
The dataset of agreement and disagreement comprises 53 episodes of agreement, 94 episodes of disagreement, and 130 neutral episodes of neither agreement or disagreement. These feature 28 participants and they occur over 11 political debates. We used automatically extracted prosodic features (continuous), and manually annotated hand and head gestures (binary). We compared the finite HCRF and the IHCRF–MCMC to our HCRF–DPM based on the F1 measure they achieved. In each case, we evaluated their performance on a test set consisting of sequences from 3 debates. We ran all models with 60 random initializations, selecting the best trained model each time by examining the F1 achieved on a validation set consisting of sequences from 3 debates. It is important to stress that each sequence belonged uniquely to either the training, the validation, or the testing set.

The database of pain we used includes 25 subjects expressing various levels of pain in 200 video sequences. Our features were based on the presence (binary) of each of the 45 observable facial muscle movements–Action Units (AUs) [15]. For our experiments, we compared the finite HCRF and the IHCRF–MCMC to our HCRF–DPM based on the F1 measure they achieved. We evaluated the

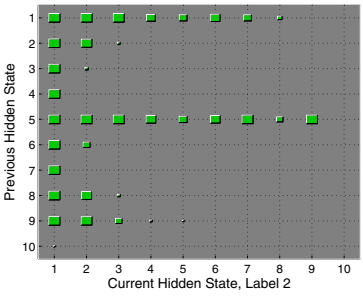
performance of the models on 25 different folds (leave–7–subjects–out for testing). In each case we concatenated the predictions for every test sequence of each fold and calculated the F1 measure for each label. The measure we used was the average F1 over all labels. We ran all experiments with 10 random initializations, selecting the best model each time by examining the F1 achieved on a validation set consisting of the sequences from 7 subjects. In every fold our training, validation and testing sets comprised not only of unique sequences but also of unique subjects.



(a) π_x (green) and π_y (black)—$L = 10$



(b) π_e, Label 1—$L = 10$



(c) π_e, Label 2—$L = 10$

Fig. 3. Hinton Diagrams of π-quantities in node and edge features of variational HCRF-DPM models with truncation level $L = 10$ for ADA2. The first column presents the π-quantities for node features: π_x for observation features in green, π_y for labels in black. The second and third columns present the π_e-quantities for labels 1 and 2 respectively. See text for additional details.

For all four tasks, in addition to the random initializations the best HCRF model was also selected by experimenting with different number of hidden states and different values for the HCRF L2 regularization coefficient. Specifically, for each random initialization we considered models with 2, 3, 4, and 5 hidden states and an L2 coefficient of 1, 10, and 100. This set of values for the hidden states was selected after preliminary results deemed a larger number

of hidden states only resulted in severe overfitting for all problems. We did not use regularization for our HCRF-DPM models and all of them had their truncation level set to $L = 10$ and their hyperparameters to $s_1 = 1000$ and $s_2 = 10$. Finally, our finite HCRF models were trained with a maximum of 300 iterations for the gradient ascent method used [10], whereas our HCRF-DPM models were trained with a maximum of 1200 variational coordinate descent iterations and a maximum of 600 iterations of gradient descent. All IHCRF–MCMC models were trained according to the experimental protocol of [7]. They had their initial number of represented hidden states set to K = 10, they were trained with 100 sampling iterations, and were tested by considering 100 samples.

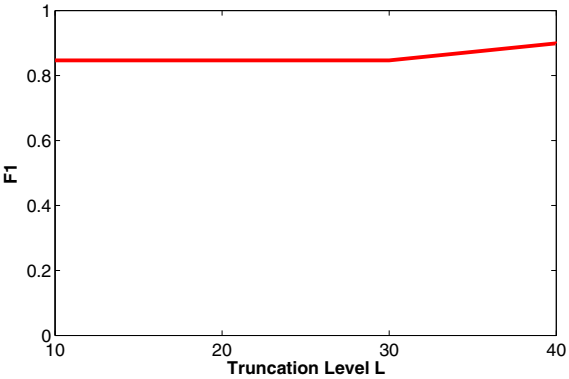


Fig. 4. HCRF–DPM F1 measure (higher F1 means higher perfomance) achieved on the validation set of ADA2. Our model does not show signs of overfitting: the F1 achieved on the validation set does not decrease as the truncation level L, and thus the number of $\boldsymbol{\theta}$–parameters, increases.

In Fig. 3 we show the learned nonparametric π parts of the features of the best HCRF–DPM ADA2 model, based on F1 achieved on our validation set, for truncation level $L = 10$. Each row is a separate DPM; with the DPMs for the edge potentials spanning across labels. Recall from Fig. 2 that these quantities have to sum to 1 across each row. As one can see in these figures, setting the truncation level $L = 10$ was a reasonable choice. Paying particular attention to the first column (node features), it seems that HCRF–DPMs converge to a small number of utilized hidden states —the equivalent table for a finite HCRF would be dense with each state being used. One can see unique and shared states, a feature of HCRFs that makes them particularly appealing for classification tasks. Fig. 3a clearly shows that the model uses only two states, one of them (state 1) being shared among both labels –features 12 and 13 in this Hinton diagram– and another (state 5) being used only by label 2.

Since we have introduced parameters $\boldsymbol{\theta}$ it is sensible to test our methodology for signs of overfitting. The only value linked with the number of our parameters

is our truncation level L: their number increases as we increase L. In Fig. 4 we show the F1 measure achieved on the validation set of ADA2 for HCRF–DPMs with L=10, 20, 30, 40. This graph is a strong indication that HCRF–DPMs do not show signs of overfitting. We would see such signs if by increasing L the performance (F1 measure) for our validation set would decrease. However, as we see here, performance on the validation sets remains roughly the same as we increase L.

Table 1. F1 measure achieved by our HCRF-DPM vs. the best, in each fold of each problem, finite HCRF and IHCRF-MCMC. **Synthetic:** Two–label classification for an HMM–generated dataset with continuous–only features **ADA2:** Two–label classification for the Canal9 Dataset of agreement and disagreement; **ADA3:** Three-label classification for the Canal9 Dataset; **PAIN2:** Two–label classification for the UNBC dataset of shoulder pain; **PAIN3:** Three–label classification for the UNBC dataset

Dataset	Finite HCRF	IHCRF–MCMC	Our HCRF–DPMs
Synthetic	100.0%	33.3%	**100.0%**
ADA2	58.4%	61.2%	**76.1**%
ADA3	50.7%	**60.3**%	49.8%
PAIN2	83.9%	88.4%	**89.2**%
PAIN3	53.9%	57.7%	**59.0**%

Table 1 shows the average over all labels of the F1 measure on the test sets for all our problems. Since the nonparametric model structure is not specified a priori but is instead determined from our data, the HCRF–DPM model is more flexible than the finite HCRF and is able to achieve better performance in all cases, with the exception of 3-label classification problem of agreement/disagreement (ADA3), where the HCRF–DPM seems to perform almost equally well with the finite model. The HCRF–DPM perfomed better than the IHCRF–MCMC in all problems with the exception of ADA3. An analysis of a IHCRF–MCMC model trained for ADA3 shows that the model ignored the two continuous dimensions and used only the binary features to model the dataset, which evidently resulted in slightly better performance.

5 Conclusion

In this paper we have presented a novel variational approach to learning an infinite Hidden Conditional Random Field, the HCRF–DPM, a discriminative nonparametric sequential model with latent variables. This deterministic approach overcomes the limitations of sampling techniques, like the one presented in [7]. We have also shown that our model is in fact a generalization of the IHCRF–MCMC presented in [7] and is able to handle sequence classification problems with continuous features naturally. In support of the latter claim, we conducted an experiment with a Gaussian HMM–generated synthetic dataset of

continuous–only features which showed that HCRF–DPMs are able to perform well on classification problems where the IHCRF–MCMC fails. Furthermore, we conducted experiments with four challenging tasks of classification of naturalistic human behavior. HCRF–DPMs were able to find a good number of shared hidden states, and to perform well in all problems, without showing signs of overfitting.

Acknowledgements. This work has been funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). K. Bousmalis is a recipient of the Google Europe Fellowship in Social Signal Processing, and this research is supported in part by this Google Fellowship. The work of Maja Pantic is funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army Research, Development, and Engineering Command. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government.

References

1. Rasmussen, C.: The Infinite Gaussian Mixture Model. In: Proc. Advances in Neural Information Processing Systems, pp. 554–560 (2000)
2. Beal, M., Ghahramani, Z., Rasmussen, C.: The Infinite Hidden Markov Model. In: Proc. Advances in Neural Information Processing Systems, pp. 577–584 (2002)
3. Fox, E., Sudderth, E., Jordan, M., Willsky, A.S.: An HDP–HMM for systems with state persistence. In: Proc. Int'l Conf. on Machine Learning (2008)
4. Orbanz, P., Buhmann, J.: Nonparametric Bayes Image Segmentation. Int'l Journal in Computer Vision 77, 25–45 (2008)
5. Van Gael, J., Teh, Y., Ghahramani, Z.: The Infinite Factorial Hidden Markov Model. Advances in Neural Information Processing Systems 21, 1697–1704 (2009)
6. Chatzis, S., Tsechpenakis, G.: The infinite hidden Markov random field model. IEEE Trans. Neural Networks 21(6), 1004–1014 (2010)
7. Bousmalis, K., Zafeiriou, S., Morency, L.P., Pantic, M.: Infinite hidden conditional random fields for human behavior analysis. IEEE Trans. Neural Networks and Learning Systems 24(1), 170–177 (2013)
8. Ghahramani, Z., Beal, M.: Propagation Algorithms for Variational Bayesian Learning. Advances in Neural Information Processing Systems 13, 507–513 (2001)
9. Sethuraman, J.: A Constructive Definition of Dirichlet Priors. Statistica Sinica 4, 639–650 (1994)
10. Quattoni, A., Wang, S., Morency, L., Collins, M., Darrell, T.: Hidden Conditional Random Fields. IEEE Trans. Pattern Analysis and Machine Intelligence, 1848–1852 (2007)
11. Blei, D., Jordan, M.: Variational Inference for Dirichlet Process Mixtures. Bayesian Analysis 1(1), 121–144 (2006)

12. Bertsekas, D.: On the Goldstein-Levitin-Polyak Gradient Projection Method. IEEE Trans. on Automatic Control 21, 174–184 (1976)
13. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
14. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an Emerging Domain. Image and Vision Computing 27(12), 1743–1759 (2009)
15. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System. Research Nexus, Salt Lake City (2002)